postfix pattern "* builds", and the weight 1.2 assigned to "Foobar" is the sum of the two patterns' weights, 0.7 and 0.5. The other identified term "cars" has a weight of 0.8 because the matching prefix pattern "world's best *" has a weight of 0.8. In some embodiments the weight for each term is computed using a log transform, where the final weight is equal to log(initial weight +1). It is possible that the two terms "Foobar" and "cars" may not be in the training data **750** and may have never been encountered by the user before. Nevertheless, the context analysis method described above identifies these terms and adds them to the user's term-based profile. Thus, context analysis can be used to discover terms associated with a particular documents, where the documents are those associated with the user, and thus the user's interests and preferences.

[**0072**] As noted, the output of context analysis can be used directly in constructing a user's term-based profile. Additionally, it may be useful in building other types of user profiles, such as a user's category-based profile. For example, a set of weighted terms can be analyzed and classified into a plurality of categories covering different topics, and those categories can be added to a user's category-based profile.

[**0073**] After executing the context analysis on a set of documents identified by or for a user, the resulting set of terms and weights may occupy a larger amount of storage than allocated for each user's term-based profile. Also, the set of terms and corresponding weights may include some terms with weights much, much smaller than other terms within the set. Therefore, in some embodiments, at the conclusion of the context analysis, the set of terms and weights is pruned by removing terms having the lowest weights (A) so that the total amount of storage occupied by the term-based profile meets predefined limits, and/or (B) so as to remove terms whose weights are so low, or terms that correspond to older items, as defined by predefined criteria, that the terms are deemed to be not indicative of the user's search preferences and interests. In some embodiments, similar pruning criteria and techniques are also applied to the category-based profile and/or the link-based profile.

[**0074**] In some embodiments, a user's profile is updated in the above manner each time the user performs a search and selects at least one document from the search results to download or view. In some embodiments, the personalization server **108** builds a list of documents identified by the user (e.g., by selecting the documents from search results) over time, and at predefined times (e.g., when the list reaches a predefined length, or a predefined amount of time has elapsed), performs a profile update of the user profile. When performing an update, new profile data is generated, and the new profile data is merged with the previously generated profile data for the user. In some embodiments, the new profile data is assigned higher importance than the previously generated profile data, thereby enabling the system to quickly adjust a user's profile in accordance with changes in the user's search preferences and interests. For example, the weights of items in the previously generated profile data may be automatically scaled downward prior to merging with the new profile data. In one embodiment, there is a date associated with each item in the profile, and the information in the profile is weighted based on its age, with older items receiving a lower weight than when they were new. In other embodiments, the new profile data is not assigned high importance than the previously generated profile data.

[**0075**] The paragraph sampling and context analysis methods may be used independently or in combination. When used in combination, the output of the paragraph sampling is used as input to the context analysis method. When used alone, the context analysis method can take the entire text of a document as its input, rather than just a sample.

[**0076**] Personalization of Search Results with the User Profile

[**0077**] The above-described methods used for creating user profiles, e.g., paragraph sampling and context analysis, may be also leveraged for determining the relevance of a candidate document to a user's preference, and thereby personalizing the results of a given search. Indeed, one function of the system **100** is to identify a set of documents that are most relevant to a user's interests based on both the user's search query as well as the user's user profile. **FIG. 8** illustrates several exemplary data structures that can be used to store information about a document's relevance to a user profile from multiple perspectives. As noted above, the search engine **104** retrieves a set of documents that form the search results. These documents are herein called "candidate documents", since they are candidates that may be potentially provided to the user. For each candidate document, identified by a respective DOC_ID, term-based document information table **810** includes multiple pairs of terms and their weights, category-based document information table **830** includes a plurality of categories and associated weights, and link-based document information table **850** includes a set of links and corresponding weights.

[**0078**] The rightmost column of each of the three tables (**810**, **830** and **850**) stores the rank (or a computed score) of a document when the document is evaluated using the particular type of user profile associated with the table. A user profile rank for a given document can be determined by combining the weights of the items (columns) associated with a document. For instance, a category-based or topic-based profile rank may be computed as follows. A user may prefer documents associated with the "Science" category with a weight of 0.6, while he dislikes documents about the "Business" category with a weight of −0.2. Thus, when a document that is within the "Science" category matches a search query, it will be weighted higher than a document in the "Business" category. In general, the document topic classification may not be exclusive. A candidate document may be classified as being a science document with probability of 0.8 and a business document with probability of 0.4. A link-based profile rank may be computed based on the relative weights allocated to a user's URL, host, domain, etc., preferences in the link-based profile. In one embodiment, term-based profile rank can be determined using known techniques, such as the term frequency-inverse document frequency (TF-IDF). The term frequency of a term is a function of the number of times the term appears in a document. The inverse document frequency is an inverse function of the number of documents in which the term appears within a collection of documents. For example, very common terms like "the" occur in many documents and consequently as assigned a relatively low inverse document frequency.